

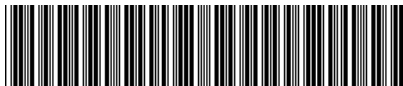


100029

北京市朝阳区安定路 35 号六层 35—10—2 内 620 室
北京弘权知识产权代理有限公司 逯长明,孙亚芹(010-84195899)

发文日:

2025 年 05 月 16 日



申请号或专利号: 202410022841.6

发文序号: 2025051301987560

案件编号: 4W118569

发明创造名称: 基于音频生成动态图像的方法、装置、设备及存储介质

专利权人: 南京硅基智能科技集团股份有限公司

无效宣告请求人: 深圳市趣推科技有限公司

无效宣告请求审查决定书

(第 584244 号)

根据专利法第 46 条第 1 款的规定,国家知识产权局对无效宣告请求人就上述专利权所提出的无效宣告请求进行了审查,现决定如下:

☐宣告专利权全部无效。

☐宣告专利权部分无效。

☒维持专利权有效。

根据专利法第 46 条第 2 款的规定,对本决定不服的,可以在收到本通知之日起 3 个月内向北京知识产权法院起诉,对方当事人作为第三人参加诉讼。

附:决定正文 17 页(正文自第 2 页起算)。

合议组组长:熊婷

主审员:吴士芬

参审员:林甦



201019
2022.10

纸件申请,回函请寄:100088 北京市海淀区蓟门桥西土城路 6 号 国家知识产权局专利局复审和无效审理部收
电子申请,应当通过专利业务办理系统以电子文件形式提交相关文件。除另有规定外,以纸件等其他形式提交的文件视为未提交。

国家知识产权局

无效宣告请求审查决定(第 584244 号)

案件编号	第 4W118569 号
决定日	2025 年 04 月 26 日
发明创造名称	基于音频生成动态图像的方法、装置、设备及存储介质
国际主分类号	G06T 13/40
无效宣告请求人	深圳市趣推科技有限公司
专利权人	南京硅基智能科技有限公司
专利号	202410022841.6
申请日	2024 年 01 月 08 日
授权公告日	2024 年 05 月 07 日
无效宣告请求日	2024 年 08 月 01 日
法律依据	专利法第 22 条第 3 款
决定要点： 涉及人工智能应用的技术方案通常包括模型设计、模型训练、模型应用三个阶段。面对涉及人工智能应用的权利要求时，应站位本领域技术人员，从处理目的、处理对象、处理结果、结果使用等方面准确理解权利要求的技术术语，确定权利要求记载的技术手段的所处阶段，进而准确地判断本专利和现有技术记载的技术方案的异同。 模型训练阶段是涉及人工智能应用的技术方案中特定而重要的部分。如果涉及人工智能应用的权利要求的技术方案与最接近的现有技术虽用于相同的场景但采用了不同的模型训练方法，现有技术整体上未给出改进模型训练方法的启示，且上述不同的模型训练方法使得该权利要求的技术方案相对于现有技术获得了诸如模型输出结果更精确或模型训练成本更低的有益的技术效果，则权利要求的技术方案具备创造性。	

一、案由

本专利的专利号为 202410022841.6，申请日为 2024 年 01 月 08 日，授权公告日为 2024 年 05 月 07 日。本专利授权公告时的权利要求书如下：

“1.一种基于音频生成动态图像的方法，其特征在于，包括：

获取用户输入的参考图像和参考音频；

基于所述参考图像和训练后的生成网络模型，确定目标头部动作特征和目标表情系数特征；所述训练后的生成网络模型用于基于输入的所述参考图像生成多个预测图像，并基于各所述预测图像和所述参考图像之间的区别，确定目标头部动作特征和目标表情系数特征；

基于所述目标头部动作特征和所述目标表情系数特征对所述训练后的生成网络模型进行调整，得到目标生成网络模型；

基于所述参考音频、所述参考图像和所述目标生成网络模型，对待处理图像进行处理，得到目标动态图像；所述目标动态图像表征所述待处理图像中目标人物基于所述参考音频变化面部表情的动态图像；所述待处理图像与所述参考图像中的图像对象相同；所述目标生成网络模型用于基于输入的所述参考音频和所述参考图像得到目标驱动特征，并基于所述目标驱动特征对输入的所述待处理图像中的目标区域进行驱动以输出所述目标动态图像。

2.根据权利要求 1 所述的方法，其特征在于，所述基于所述参考图像和训练后的生成网络模型，确定目标头部动作特征和目标表情系数特征，包括：

基于所述参考图像生成的预设时长的视频以得到参考数据；所述参考数据通过对基于参考图像进行处理得到；

通过所述训练后的生成网络模型从所述参考数据中提取所述目标头部动作特征和所述目标表情系数特征。

3.根据权利要求 2 所述的方法，其特征在于，所述通过所述训练后的生成网络模型从所述参考数据中提取所述目标头部动作特征和所述目标表情系数特征，包括：

通过所述训练后的生成网络模型从所述参考数据中确定多帧目标人物的面部图像；

通过所述训练后的生成网络模型基于各帧所述目标人物的面部图像进行特征提取，得到所述目标头部动作特征和所述目标表情系数特征。

4.根据权利要求 1 所述的方法，其特征在于，所述目标生成网络模型包括仿射子网络和驱动子网络；

所述基于所述参考音频、所述参考图像和所述目标生成网络模型，对待处理图像进行处理，得到目标动态图像，包括：

通过所述仿射子网络对所述待处理图像进行处理，得到待处理特征图，并通过所述仿射子网络基于所述参考音频、所述参考图像和所述待处理特征图，得到形变特征图；所述仿射子网络用于确定所述参考音频对应的目标梅尔倒谱系数特征，对所述参考图像进行特征提取得到参考特征图，以及对所述参考特征图进行仿

射变换得到所述形变特征图；

通过所述驱动子网络基于所述形变特征图，对所述待处理图像进行处理，得到所述目标动态图像；所述驱动子网络用于对所述待处理图像进行驱动处理，得到所述目标动态图像。

5.根据权利要求4所述的方法，其特征在于，所述仿射子网络包括语音处理层、特征提取层、特征融合层和特征仿射层；

所述通过所述仿射子网络对所述待处理图像进行处理，得到待处理特征图，并通过所述仿射子网络基于所述参考音频、所述参考图像和所述待处理特征图，得到形变特征图，包括：

通过所述语音处理层确定所述参考音频对应的目标梅尔倒谱系数特征；

通过所述特征提取层对所述参考图像进行特征提取，得到参考特征图；

通过所述特征提取层对所述待处理图像进行特征提取，得到待处理特征图；

通过所述特征融合层对所述参考特征图和所述待处理特征图进行特征堆叠对齐处理，得到融合特征图；

通过所述特征仿射层基于所述融合特征图和所述目标梅尔倒谱系数特征确定仿射系数，并基于所述仿射系数对所述参考特征图进行仿射变换的空间形变，得到所述形变特征图。

6.根据权利要求5所述的方法，其特征在于，所述驱动子网络包括特征处理层、特征同步层和图像驱动层；

所述通过所述驱动子网络基于所述形变特征图，对所述待处理图像进行处理，得到所述目标动态图像，包括：

通过所述特征处理层基于所述目标梅尔倒谱系数特征，得到初始驱动特征；

通过所述图像驱动层基于所述初始驱动特征对所述待处理特征图进行驱动处理，得到初始特征图；

通过所述特征同步层对所述形变特征图与所述初始特征图进行堆叠处理，以确定所述形变特征图与所述初始特征图之间的特征同步参数；

通过所述特征处理层基于所述特征同步参数对所述初始驱动特征进行调整，得到目标驱动特征；

通过所述图像驱动层基于所述目标驱动特征，对所述待处理图像进行驱动处理，得到所述目标动态图像。

7.根据权利要求1-6任一项所述的方法，其特征在于，所述方法还包括：

获取样本视频；其中，所述样本视频中的视频对象与所述待处理图像中的图像对象不同；

通过待训练的生成网络模型对所述样本视频进行处理，提取样本音频数据和样本图像数据；

基于所述待训练的生成网络模型对所述样本音频数据和所述样本图像数据进行处理，得到预测训练结果；

以所述预测训练结果作为所述待训练的生成网络模型的初始训练输出，所述样本图像数据作为监督信息，迭代训练所述待训练的生成网络模型得到所述训练后的生成网络模型。

8.根据权利要求7所述的方法，其特征在于，所述基于所述待训练的生成网络模型对所述样本音频数据和所述样本图像数据进行处理，得到预测训练结果，包括：

基于所述待训练的生成网络模型从所述样本音频数据中提取参考梅尔倒谱系数特征；

基于所述待训练的生成网络模型从所述样本图像数据中提取参考头部动作、参考表情系数特征和参考人脸特征；

通过所述待训练的生成网络模型基于所述参考梅尔倒谱系数特征、所述参考头部动作、所述参考表情系数特征和所述参考人脸特征，得到所述预测训练结果。

9.根据权利要求 7 所述的方法，其特征在于，所述以所述预测训练结果作为所述待训练的生成网络模型的初始训练输出，所述样本图像数据作为监督信息，迭代训练所述待训练的生成网络模型得到所述训练后的生成网络模型，包括：

根据所述预测训练结果和所述样本图像数据，确定损失值；

根据所述损失值，迭代更新所述待训练的生成网络模型，得到所述训练后的生成网络模型。

10.一种基于音频生成动态图像的装置，其特征在于，包括：

获取模块，用于获取用户输入的参考图像和参考音频；

处理模块，用于基于所述参考图像和训练后的生成网络模型，确定目标头部动作特征和目标表情系数特征；所述训练后的生成网络模型用于基于输入的所述参考图像生成多个预测图像，并基于各所述预测图像和所述参考图像之间的区别，确定目标头部动作特征和目标表情系数特征；

调整模块，用于基于所述目标头部动作特征和所述目标表情系数特征对所述训练后的生成网络模型进行调整，得到目标生成网络模型；

所述处理模块，还用于基于所述参考音频、所述参考图像和所述目标生成网络模型，对待处理图像进行处理，得到目标动态图像；所述目标动态图像表征所述待处理图像中目标人物基于所述参考音频变化面部表情的动态图像；所述待处理图像与所述参考图像中的图像对象相同；所述目标生成网络模型用于基于输入的所述参考音频和所述参考图像得到目标驱动特征，并基于所述目标驱动特征对输入的所述待处理图像中的目标区域进行驱动以输出所述目标动态图像。

11.一种电子设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求 1-9 任一项所述的方法的步骤。

12.一种非暂态计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求 1-9 任一项所述的方法的步骤。”

请求人于 2024 年 08 月 01 日向国家知识产权局提出了无效宣告请求，其理由是权利要求 1-12 不符合专利法第 22 条第 3 款有关创造性的规定，请求宣告本专利权利要求 1-12 全部无效，同时提交了如下证据：

证据 1：CSDN 博客网页，胖胖腐乳，“SadTalker Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking”，https://blog.csdn.net/weixin_41967328/article/details/129517368，发布日期为 2023 年 03 月 14 日，及其中文译文；

证据 2: CSDN 博客网页, Redflashing, “[读论文]弱监督学习的精确 3D 人脸重建: 从单个图像到图像集—Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set”, https://blog.csdn.net/weixin_42363722/article/details/120254880, 发布日为 2021 年 09 月 12 日, 及其中文译文;

证据 3: 申请公布号为 CN115131849A 的中国发明专利申请, 申请公布日为 2022 年 09 月 30 日;

证据 4: 申请公布号为 CN117315157A 的中国发明专利申请, 申请公布日为 2023 年 12 月 29 日;

证据 5: 申请公布号为 CN116912639A 的中国发明专利申请, 申请公布日为 2023 年 10 月 20 日;

证据 6: 申请公布号为 CN116844215A 的中国发明专利申请, 申请公布日为 2023 年 10 月 03 日;

证据 7: 申请公布号为 CN115035604A 的中国发明专利申请, 申请公布日为 2022 年 09 月 09 日。

请求人主张: (1) 以证据 1 作为最接近的现有技术, 权利要求 1、10 相对于证据 1 和证据 2 的结合, 或证据 1 和公知常识的结合不具备创造性; 权利要求 2-3 相对于证据 1、证据 2 和证据 7 的结合, 或证据 1 和公知常识的结合不具备创造性; 权利要求 4-9 相对于证据 1、证据 2、证据 6、证据 7 和公知常识的结合, 或证据 1、证据 6 和公知常识的结合不具备创造性; 基于对权利要求 1-9 的引用关系, 权利要求 11-12 相对于相应证据组合也不具备创造性; (2) 以证据 3 作为最接近的现有技术, 权利要求 1-3、10 相对于证据 3 和证据 4 的结合, 或证据 3 的各个实施例结合不具备创造性; 权利要求 4-9 相对于证据 3、证据 4、证据 6 和公知常识的结合, 或证据 3、证据 6 和公知常识的结合不具备创造性; 基于对权利要求 1-9 的引用关系, 权利要求 11-12 相对于相应证据组合也不具备创造性; (2) 以证据 5 作为最接近的现有技术, 权利要求 1 相对于证据 5 不具备创造性; 权利要求 2 相对于证据 5 和公知常识的结合不具备创造性。

经形式审查合格, 国家知识产权局于 2024 年 08 月 12 日受理了上述无效宣告请求并将无效宣告请求书及证据副本转给了专利权人, 同时成立合议组对本案进行审查。

国家知识产权局本案合议组于 2024 年 10 月 17 日向双方当事人发出了口头审理通知书, 定于 2024 年 12 月 20 日举行口头审理,

专利权人针对上述无效宣告请求于 2024 年 12 月 18 日提交了意见陈述书, 并认为请求人的证据均不涉及本专利权利要求 1 中的模型微调过程, 请求人的无效理由均不成立。

本案合议组于 2024 年 12 月 19 日将上述意见陈述书转送给请求人。

口头审理如期举行, 双方当事人均委托代理人出席了本次口头审理。在口头审理过程中, 双方当事人对方出庭人员的身份无异议, 对合议组成员变更无异议, 对合议组成员和书记员无回避请求, 并明确了以下事项:

(1) 请求人明确其无效理由、证据和范围以请求书为准; 涉及权利要求 1 的创造性最主要的结合方式为证据 1 结合证据 2; 权利要求 1 相对于证据 5 的多个实施例结合不具备创造性的理由并不涉及公知常识;

(2) 专利权人对请求人提交的证据 1-7 的真实性、公开日期无异议, 对证据 1、证据 2 的中文译文的准确性无异议; 并明确其于 2024 年 12 月 18 日提交的意见陈述仅供合议组参考;

(3) 双方当事人已经充分陈述了意见，口头审理之后合议组不再接受双方当事人的意见陈述。

至此，合议组认为本案事实已经清楚，可以作出审查决定。

二、决定的理由

(一) 审查基础

本无效宣告请求决定的审查基础为本专利的授权公告文本。

(二) 证据认定

证据 1 和证据 2 为 CSDN 网站的博客网页，证据 3—证据 7 为中国专利文献，专利权人对请求人提交的证据 1—7 的真实性、公开日期无异议，对证据 1、证据 2 的中文译文的准确性无异议；合议组亦未发现影响证据 1—7 的真实性的明显瑕疵，因此对证据 1—7 的真实性、公开日期予以确认。证据 1—7 的公开日期均在本专利的申请日之前，属于可以用于评述本专利是否符合专利法第 22 条第 3 款规定的创造性的现有技术。证据 1、证据 2 的中文译文以请求人提交的中文译文为准。

(三) 关于权利要求技术方案的理解

涉及人工智能应用的技术方案通常包括模型设计、模型训练、模型应用三个阶段。面对涉及人工智能应用的权利要求时，应站位本领域技术人员，从处理目的、处理对象、处理结果、结果使用等方面准确理解权利要求的技术术语，确定权利要求记载的技术手段的所处阶段，进而准确地判断本专利和现有技术记载的技术方案的异同。

权利要求 1 要求保护 (1-1) 一种基于音频生成动态图像的方法，其特征在于，包括：

(1-2) 获取用户输入的参考图像和参考音频；

(1-3) 基于所述参考图像和训练后的生成网络模型，确定目标头部动作特征和目标表情系数特征；所述训练后的生成网络模型用于基于输入的所述参考图像生成多个预测图像，并基于各所述预测图像和所述参考图像之间的区别，确定目标头部动作特征和目标表情系数特征；

(1-4) 基于所述目标头部动作特征和所述目标表情系数特征对所述训练后的生成网络模型进行调整，得到目标生成网络模型；

(1-5) 基于所述参考音频、所述参考图像和所述目标生成网络模型，对待处理图像进行处理，得到目标动态图像；所述目标动态图像表征所述待处理图像中目标人物基于所述参考音频变化面部表情的动态图像；所述待处理图像与所述参考图像中的图像对象相同；所述目标生成网络模型用于基于输入的所述参考音频和所述参考图像得到目标驱动特征，并基于所述目标驱动特征对输入的所述待处理图像中的目标区域进行驱动以输出所述目标动态图像。

根据本专利权利要求 1 结合说明书第[0004]段的记载，权利要求 1 的技术方案是基于目标人物的单张图片 and 音频数据并利用神经网络模型得到目标人物的动态图像，具体技术手段包括权利要求 1 中的特征 (1-1) 至 (1-5)，其中特征 (1-3) 描述了训练后的生成网络模型的处理对象为用户输入的参考图像，以及模型生成

的结果为目标头部动作特征和目标表情系数特征，特征（1-4）描述了利用特征（1-3）生成的结果对生成网络模型进行调整；可见特征（1-3）-（1-4）描述的是对神经网络模型进行调整的过程，其属于神经网络模型的模型训练阶段。特征（1-5）描述了神经网络模型的处理对象为参考音频和参考图像以及待处理图像，模型的生成结果为目标动态图像，该生成的结果即为目标对象的动态图像，属于模型应用阶段。

（四）关于专利法第 22 条第 3 款

专利法第 22 条第 3 款规定：创造性，是指与现有技术相比，该发明具有突出的实质性特点和显著的进步，该实用新型具有实质性特点和进步。

模型训练阶段是涉及人工智能应用的技术方案中特定而重要的部分。如果涉及人工智能应用的权利要求的技术方案与最接近的现有技术虽用于相同的场景但采用了不同的模型训练方法，现有技术整体上未给出改进模型训练方法的启示，且上述不同的模型训练方法使得该权利要求的技术方案相对于现有技术获得了诸如模型输出结果更精确或模型训练成本更低的有益的技术效果，则权利要求的技术方案具备创造性。

（4.1）关于权利要求 1

（4.1.1）以证据 1 作为最接近的现有技术

证据 1（参见证据 1 中文译文第 1 节-第 4 节及图 1-8）：公开了一种通过隐式三维系数调制的程式化音频驱动的说话头视频生成系统，称为“SadTalker”（相当于本专利“一种基于音频生成动态图像的方法”，即公开了权利要求 1 的特征（1-1）），具体如下：如图 2 所示，我们的系统使用三维运动系数作为说话头生成的中间表示。我们首先从原始图像（相当于本专利“获取用户输入的参考图像”）中提取系数。然后利用 ExpNet 和 PoseVAE 分别生成逼真的 3DMM 运动系数。最后，提出了一种 3d 感知的面部渲染来生成说话头部视频。下面，我们将在第 3.1 节中简要介绍 3D 人脸模型，分别在第 3.2 节和第 3.3 节中设计音频驱动的运动系数生成和系数驱动的图像动画器。（如图 2 所示，输入音频相当于本专利的“参考音频”）

3.1 三维人脸模型的初步研究

受最近的单幅图像深度三维重建方法[4]的启发，我们将预测的三维变形模型（3DMMs）的空间作为我们的中间表示。在 3DMM 中，三维面形 S 可以解耦为：

$$S = \bar{S} + \alpha U_{id} + \beta U_{exp},$$

其中 $\bar{S}$ 为 3D 人脸的平均形状， U_{id} 和 U_{exp} 为 LSF 的变形模型[1]的身份和表达的标准正交基。系数 $\alpha \in \mathbb{R}^{80}$ 和 $\beta \in \mathbb{R}^{64}$ 分别描述了人的身份和表达，为了保持位姿方差， $r \in SO(3)$ 和 $t \in \mathbb{R}^3$ 表示头部旋转和平移。为了实现同一性无系数生成[31]，我们只对运动参数建模为 $\{\beta, r, t\}$ 。我们从前面介绍的驾驶音频中分别学习了头部姿势 $p = [r, t]$ 和表达系数 β 。然后，这些运动系数用于隐式调节我们的面部渲染，以实现最终的视频合成。

3.2 通过音频生成运动系数

如上所述，3D 运动系数包含头部姿态和表情，其中头部姿态是全局运动，而表情是相对局部的。因为头部的姿势与音频的关系较弱，而嘴唇的运动是高度相关。我们使用提出的 PoseVAE 和 ExpNet 生成头部姿态

和表情的运动（相当于本专利的“目标驱动特征”），分别介绍如下。

对于身份问题，我们通过第一帧的表达系数 β_0 将表情运动与特定的人联系起来。为了减少自然说话中其他面部成分的运动权重，我们通过 Wav2Lip[28]预训练网络和深度三维重建[4]，只使用嘴唇运动系数作为系数目标。然后，其他微小的面部动作(例如眨眼)可以通过渲染图像上的额外地标损失来利用。

图 3。我们的 ExpNet 结构。我们采用单目三维人脸重建模型[4]（ R_e 和 R_d ）来学习真实表情系数。其中 R_e 是预训练的 3DMM 系数估计器， R_d 是没有可学习参数的可微分 3D 人脸渲染。我们使用参考表达式 β_0 为来降低身份的不确定性，并使用预训练的 Wav2Lip[28]生成的帧和第一帧作为目标表达系数，因为它只包含与嘴唇相关的运动。

图 4。PoseVAE 管道。我们通过条件 VAE 结构学习输入头位 ρ_0 的残差。给定条件：第一帧 ρ_0 ，样式标识 Z_{style} 和音频剪辑 $\alpha\{1, ..., t\}$ 时，我们的方法学习了残差头位姿 $\Delta\rho\{1, ..., t\} = \rho\{1, ..., t\} - \rho_0$ 。训练之后，我们可以只通过姿态解码器和条件（cond.）生成程式化的结果。

如图 4 所示，我们设计了一个基于 VAE[20]的模型来学习真实说话视频中真实的、身份感知的程式化头部运动 $\rho \in \mathbb{R}^6$ 。在训练中，姿态 VAE 使用基于编码器的结构在固定 n 帧上进行训练。编码器和解码器都是两层 MLP，其中输入包含一个连续的 t 帧头姿态，我们将其嵌入到高斯分布中。在解码器中，学习网络从采样分布中生成 t 帧姿态。我们的 PoseVAE 不是直接生成姿态，而是学习第一帧的条件姿态 ρ_0 的残差，这使得我们的方法能够在第一帧条件下生成更长的、稳定的、连续的头部运动。另外，根据 CVAE[5]，我们添加相应的音频特征 $\alpha\{1, ..., t\}$ 和风格认同 Z_{style} 作为节奏意识和风格认同的条件。采用 kl-散度 $\mathcal{L}_{KL}$ 测量所产生运动的分布。采用均方损耗 $\mathcal{L}_{MSE}$ 和对抗损耗 $\mathcal{L}_{GAN}$ 来保证生成的质量。我们在补充材料中提供了更详细的损失函数。

3.3 3d 感知面部渲染

图 5。提出的 FaceRender 及与 facevid2vid[40]的比较。给定源图像 I_s 和驱动图像 I_d ，facevid2vid 在无监督的三维关键点空间 X_c 、 X_s 和 X_d 中生成运动。然后通过外观 A_0 的关键点生成图像。在生成逼真的 3D 运动系数后，我们通过一个设计良好的 3D 感知图像动画器渲染最终的视频。（相当于本专利的“所述目标动态图像表征所述待处理图像中目标人物基于所述参考音频变化面部表情的动态图像”）。

4 实验

4.1 实现细节和度量

数据集。我们使用 VoxCeleb[25]数据集进行训练，其中包含 1251 个受试者的超过 100k 个视频。我们按照前面的图像动画方法[34]裁剪原始视频，并将视频大小调整为 256×256 。预处理后，数据用于训练我们的 FaceRender。由于一些视频和音频在 VoxCeleb 中没有对齐，我们选择了 46 个科目的 1890 个对齐的视频和音频来训练我们的 PoseVAE 和 ExpNet。输入音频被降采样到 16kHz，并转换为与 Wav2lip[28]设置相同的 mel-频谱图。为了测试我们的方法，我们使用了来自 HDTF 数据集[46]的 346 个视频的前 8 秒视频（总共约 70k 帧），因为它包含高分辨率和野外说话的头部视频。这些视频也按照[34]进行裁剪和处理，并将大小调整为 256

×256 以供评估。我们使用每个视频的第一帧作为参考图像来生成视频。

实现细节。ExpNet, PoseVAE 和 FaceRender 都是单独训练的, 我们使用 Adam 优化器[19]进行所有实验。经过训练后, 我们的方法可以在没有任何人工干预的情况下以端到端方式进行推断。通过预训练的深度三维人脸重建方法[4]提取所有的 3DMM 参数。所有实验均在 8 个 A100 图形处理器上进行。ExpNet、PoseV AE 和 FaceRender 的学习速率分别为 $2e^{-5}$ 、 $1e^{-4}$ 和 $2e^{-4}$ 。从时间上考虑, ExpNet 使用连续 5 帧进行学习。PoseVAE 是通过连续 32 帧学习的。FaceRender 中的帧是逐帧生成的, 为了稳定, 系数为 5 连续帧。

4.2 与其他最先进的方法进行比较

图 6, 我们将我们的方法与几种最先进的方法进行比较, 用于生成单个图像音频驱动的说话头。我们的方法在嘴唇同步、身份保留、头部运动和图像质量方面产生了更高质量的结果。

根据证据 1 的上述记载可知, 证据 1 图 2 显示了其为了实现通过人脸图像和一段语音音频生成会说话的头部视频的发明而设计的神经网络模型架构, 第 3 节详细描述了 SadTalker 模型架构, 具体的各个模块的架构如图 3-5 所示, 属于模型设计阶段; 第 4.1 节描述了 SadTalker 模型的训练过程, 该 SadTalker 的输入数据为来自 HDFT 数据集第 346 个视频的前 8 秒视频中处理的图像的音频来生成视频, 并根据评价指标评估完成了模型的训练, 可见其处理对象为样本集视频, 根据评价指标完成模型训练, 属于模型训练阶段; 第 4.2 节描述了模型应用后的效果展示以及与其它模型的比较内容, 属于模型应用阶段。即第 4.1 节训练后的 SadTalker 模型相当于本专利权利要求 1 的“训练后的生成网络模型”, 第 4.2 节使用训练后的 SadTalker 模型根据用户输入的单图像和输入的参考音频生成视频, 最终生成的视频相当于本专利权利要求 1 的“目标动态图像; 所述目标动态图像表征所述待处理图像中目标人物基于所述参考音频变化面部表情的动态图像”。证据 1 中第 4.1 节利用样本集数据训练完成 SadTalker 模型后即可直接根据输入图像和音频生成视频, 其方法中并不涉及对训练后的 SadTalker 模型再进行模型调整的内容; 而本专利权利要求 1 首先有训练后的生成网络模型, 然后再利用用户输入的图像作为参考图像对训练后的生成网络模型再进行调整, 并且这次对于模型的调整是基于输入的参考图像, 该参考图像的对象与待处理图像的对象相同, 并不涉及音频内容, 以及本专利权利要求 1 还记载了具体的模型调整手段。由此可知, 本专利权利要求 1 与证据 1 对于模型的训练方法是不同的。

可见, 本专利权利要求 1 与证据 1 的区别技术特征在于: 基于所述参考图像和训练后的生成网络模型, 确定目标头部动作特征和目标表情系数特征; 所述训练后的生成网络模型用于基于输入的所述参考图像生成多个预测图像, 并基于各所述预测图像和所述参考图像之间的区别, 确定目标头部动作特征和目标表情系数特征; 基于所述目标头部动作特征和所述目标表情系数特征对所述训练后的生成网络模型进行调整, 得到目标生成网络模型; 基于所述参考音频、所述参考图像和所述目标生成网络模型, 对待处理图像进行处理, 得到目标动态图像; 所述待处理图像与所述参考图像中的图像对象相同。基于上述区别技术特征可以确定权利要求 1 实际解决的技术问题是如何使得模型生成的对象更贴近待处理图像中对象。

证据 2 公开了 (参见证据 2 中文译文第 1 节-第 4 节及图 1-3) 一种基于 CNN 的单图像人脸重建方法。

具体如下：

图 1. (a) 方法框架：包括一个用于端到端单图像三维重建的重建网络和一个用于基于多图像重建的置信度测量子网。在本文中，作者使用 ResNet-50 网络，通过将最后一个全连接层修改为 239 个神经元，来回归这些系数。为了简洁起见，作者把这个用于单幅图像重建的修改后的 ResNet-50 网络称为 R-Net。作者将在下一节中介绍如何训练它。

4 用于单幅图像重建的混合弱监督学习

给定一个训练 RGB 图像 I ，作者使用 R-Net 回归系数向量 x ，利用它可以通过一些简单的、可微的数学推导分析生成重建的图像 I' 。R-Net 是在没有任何真实标签的情况下进行训练的，但通过评估 I' 的混合损失，并对其进行反向传播。具体的损失函数和正则化手段参见 4.1-4.3。其中 4.1.2 标记点损失：作者还使用二维图像上的人脸标记点作为弱监督来训练网络。作者使用目前最准确的三维人脸对齐方法来检测训练图像的 68 个标记点 $\{q_n\}$ 。在训练中，作者将重建的形状的三维标记顶点投射到图像上，得到 $\{q'_n\}$ ，并通过相应的公式计算出损失。

根据证据 2 的上述记载可知，R-Net 输入的对象是单幅图像，输出的结果是包括身份系数等五个系数，通过上述输出的系数进行 3D 重建与图像渲染，即可得到对应的 3D 人脸。证据 2 的目的在于提出新的混合损失计算方法，可以通过上述损失计算方法对 R-Net 训练过程的生成结果进行验证，并反向传播，最终获得更加准确的能够处理图像的神经网络模型，可见证据 2 的 R-Net 模型是用于基于单幅图像生成 3D 人脸重建所需的中间系数，其通过训练获得了对单幅图像进行处理生成系数的符合损失函数标准的模型，然后利用该训练后的模型输出系数实现 3D 人脸的重建，即证据 2 仅公开了一种用于单幅图像的神经网络模型的训练过程，训练完成后直接应用于图像推理，其虽然属于模型训练阶段，但是其模型训练阶段并不存在对训练后的模型再进行调整的内容，没有给出要在训练后的神经网络模型应用过程中利用具有相同对象的参考图像对训练后的神经网络模型进行调整的启示。正是采用了上述区别技术特征，权利要求 1 获得了生成目标动态图像中的图像更贴近待处理图像中对象的有益的技术效果。因此，权利要求 1 相对于证据 1、证据 2 的结合具有突出的实质性特点和显著的进步，具备专利法第 22 条第 3 款有关创造性的规定。

另外，证据 1 的第 3 节的模型实现了从输入的图像和音频生成目标动态图像的功能，这与本专利的模型目的相同，证据 1 的 3.2 中的 β_0 是从输入图像生成的表情系数，用于将表情运动与特定的人联系起来， I_0 是输入 Wav2Lip 预训练网络的数据，为嘴部训练的第一帧数据，两者与权利要求 1 中的基于参考图像生成的多个预测图像完全不同；而且没有证据表明上述权利要求 1 与证据 1 的区别技术特征属于本领域的公知常识，正是采用了上述区别技术特征，权利要求 1 获得了生成目标动态图像中的图像更贴近待处理图像中对象的有益的技术效果。因此，权利要求 1 相对于证据 1 和公知常识的结合具有突出的实质性特点和显著的进步，具备专利法第 22 条第 3 款规定的创造性。

(4.1.2) 以证据 3 作为最接近的现有技术

证据 3 公开了一种图像生成方法以及相关设备。（参见说明书第 76-159 段及图 1a-1d）如图 1b 所示，该图像生成方法的具体流程可以如下：

101、获取目标对象的原始面部图像帧和待生成的目标面部图像帧对应的音频驱动信息。目标对象可以为面部姿态待调整的对象，原始面部图像帧为包含目标对象面部的图像，音频驱动信息为用于对原始面部图像帧进行面部姿态调整的音频信息，目标面部图像帧具体可以为基于音频驱动信息，对原始面部图像帧中面部姿态调整后对应的面部图像。这里所说的面部姿态具体可以指面部表情，如嘴型、眼神等对象面部信息。

102、对所述原始面部图像帧进行空间特征提取，得到所述原始面部图像帧对应的原始面部空间特征。原始面部空间特征具体可以包括原始面部图像帧对应的三维(3D, 3-dimensional)面部系数，比如具体可以包括身份信息(identity)，光影(lighting)，纹理(texture)，表情(expression)，姿态(pose)，眼神(gaze)等。根据这些面部系数可以重建出原始面部图像帧的面部，如图 1d 所示。

103、对所述音频驱动信息进行时序特征提取，得到所述目标面部图像帧对应的面部局部姿态特征。对所述音频驱动信息中的各个音频帧进行特征提取，得到各个音频帧的音频语义特征信息；基于各个音频帧的前后音频帧的音频语义特征信息，对所述各个音频帧的音频语义特征信息进行处理；将处理后的各个音频帧的音频语义特征信息进行融合，得到所述目标面部图像帧对应的面部局部姿态特征。

104、基于所述原始面部空间特征和所述面部局部姿态特征，对所述目标对象进行面部重建处理，生成所述目标面部图像帧。具体为：将所述原始面部空间特征和所述面部局部姿态特征进行融合，得到融合后面部空间特征；基于所述融合后面部空间特征，对所述目标对象进行面部重建处理，得到所述目标对象对应的参考面部图像帧；基于所述原始面部图像帧、所述融合后面部空间特征和所述参考面部图像帧，生成所述目标面部图像帧。重建后三维面部图像的纹理光影可以来自于原始面部图像帧，重建后三维面部图像的姿态表情可以音频驱动信息。

（参见说明书第 160-176 段及图 1f）通过图像生成模型，……。需要说明的是，该图像生成模型可以由多组训练数据训练而成，该图像生成模型具体可以由其他设备进行训练后，提供给该图像生成装置，或者，也可以由该图像生成装置自行进行训练。

若由该图像生成装置自行进行训练，则步骤“通过图像生成模型，对所述原始面部图像帧进行空间特征提取，得到所述原始面部图像帧对应的原始面部空间特征”之前，还可以包括：

获取训练数据，所述训练数据包括样本对象的原始面部图像帧样本、目标驱动面部图像帧样本、以及所述目标驱动面部图像帧样本对应的音频驱动信息样本；通过预设图像生成模型，对所述原始面部图像帧样本进行空间特征提取，得到所述原始面部图像帧样本对应的原始面部空间特征；对所述音频驱动信息样本进行时序特征提取，得到所述目标驱动面部图像帧样本对应的面部局部姿态特征；基于所述原始面部空间特征和所述面部局部姿态特征，对所述样本对象进行面部重建处理，得到预测驱动面部图像帧；基于所述目标驱动面部图像帧样本和所述预测驱动面部图像帧，对预设图像生成模型的参数进行调整，得到训练后的图像生成

模型。其中，目标驱动面部图像帧样本可以视为标签信息，具体可以为音频驱动信息样本对应的期望驱动面部图像帧。

（参见说明书第 235–237 段及图 1g）以训练后图像生成模型输出目标面部图像帧为例，通过训练后图像生成模型输出目标分辨率下的目标面部图像帧的过程可以如图 1g 所述。由上可知，本实施例可以获取目标对象的原始面部图像帧和待生成的目标面部图像帧对应的音频驱动信息（相当于本专利的“待处理图像和输入音频”）；对所述原始面部图像帧进行空间特征提取，得到所述原始面部图像帧对应的原始面部空间特征；对所述音频驱动信息进行时序特征提取，得到所述目标面部图像帧对应的面部局部姿态特征；基于所述原始面部空间特征和所述面部局部姿态特征，对所述目标对象进行面部重建处理，生成所述目标面部图像帧。本申请可以通过对音频驱动信息进行特征提取，捕捉到目标对象部分的面部姿态细节信息，进而基于捕捉到的信息来对原始面部图像帧进行面部调整，从而获取音频驱动信息对应的目标面部图像帧，这样有利于提高目标面部图像帧的生成效率和准确性。

根据证据 3 的上述记载可知，证据 3 的图 1b 及其文字描述为一种图像生成方法，说明书第 160–165 段记载了通过“图像生成模型”完成上述方法，公开了根据图像和音频数据生成视频的模型的具体架构，属于模型设计阶段；说明书第 168–176 段为对该图像生成模型的训练阶段，其处理对象为训练模型和验证模型的样本数据，最后得到训练后的图像生成模型，其属于模型训练阶段；图 1g 及其相关文字描述为该图像生成模型具体用于根据某原始面部图像和音频驱动信息生成目标面部图像帧的过程，其属于模型应用阶段。可见证据 3 的图 1g 训练完成后的模型相当于本专利权利要求 1 的“训练后的生成网络模型”，证据 3 直接使用训练完成后的模型根据用户输入的图像和输入的参考音频生成视频，最终生成的视频相当于本专利权利要求 1 的“目标动态图像，所述目标动态图像表征所述待处理图像中目标人物基于所述参考音频变化面部表情的动态图像”，用户输入的原始面部图像相当于本专利权利要求 1 “待处理图像”。而本专利权利要求 1 首先有训练后的生成网络模型，然后再利用用户输入的图像作为参考图像对训练后的生成网络模型再进行调整，并且这次对于模型的调整是基于输入的参考图像，该参考图像的对象与待处理图像的对象相同，并不涉及音频内容，以及权利要求 1 还记载了具体的模型调整手段。由此可知，本专利权利要求 1 与证据 3 对于模型的训练方法是不同的。

可见，本专利权利要求 1 与证据 3 的区别技术特征在于：基于所述参考图像和训练后的生成网络模型，确定目标头部动作特征和目标表情系数特征；所述训练后的生成网络模型用于基于输入的所述参考图像生成多个预测图像，并基于各所述预测图像和所述参考图像之间的区别，确定目标头部动作特征和目标表情系数特征；基于所述目标头部动作特征和所述目标表情系数特征对所述训练后的生成网络模型进行调整，得到目标生成网络模型；基于所述参考音频、所述参考图像和所述目标生成网络模型，对待处理图像进行处理，得到目标动态图像；所述待处理图像与所述参考图像中的图像对象相同。基于上述区别技术特征可以确定权利要求 1 实际解决的技术问题是如何使得模型生成的对象更贴近待处理图像中对象。

证据 4 公开了一种基于单幅图像进行 3D 人脸重建的方法，并具体公开了：（参见说明书第 4 段）现有的专利都是直接使用一个神经网络同时计算出身份和表情系数，就会导致身份和表情系数存在相互耦合的问题，即不同的身份系数通过特定的表情组合，也可能能够表达出同样形状的 3D 人脸，这种现象就会造成现有方法重建出的人脸身份特征存在误差。（参见说明书第 17-35 段及图 1-2）一种基于单幅图像进行 3D 人脸重建的方法，包括如下步骤：步骤一、使用一个现有的人脸检测器对图像进行人脸检测，然后根据人脸检测的结果计算出图像中人脸区域的包围盒，并根据包围盒对图像进行裁剪，再通过 PyTorch 框架中的 Transform 模块对图像进行归一化；步骤二、使用一个现有的人脸属性编辑模型对原图像进行编辑，包括去除表情、随机产生表情以及随机旋转，在模型示意图中，(b)部分的 I_0 表示原图， I_1 表示去除表情后的图像；(c)部分的 I_2 表示随机旋转后的图像， I_3 表示随机产生表情后的图像。步骤三、3D 人脸重建的回归模型包括 ID-Net、Perform-Net 两个 ResNet50 神经网络模型，ID-Net 用于输出身份系数，身份系数是一个低维向量，表示 3D 人脸的固有特征，与表情无关，Perform-Net 用于输出表情、颜色、位姿、光照等系数；使用 3DMM 模型作为解码器。步骤四、使用步骤二中得到的去除表情后的图像与原图像配对，训练 ID-Net 和 Perform-Net 两个模型，损失函数包括特征点重投影损失、像素损失、人脸识别相似度损失、形状一致性损失以及正则损失；步骤五、将 ID-Net 网络参数固定，使用步骤二中得到的随机表情与随机旋转的图像增强结果与原图像配对，来对 Perform-Net 进一步训练调优，以增强算法针对连续视频帧表情和位姿系数提取的稳定性，损失函数包括特征点重投影损失、像素损失、位姿一致性损失、旋转一致性损失以及正则损失。（参见说明书第 12 段）本发明的有益效果是：本发明能够从野外图像中重建出具有身份信息三维人脸模型，实现了三维人脸重建的高精度、高质量和高鲁棒性。将三维人脸模型的身份系数和表情系数分离，实现了三维人脸重建的身份-表情解耦，提高了重建出的三维人脸模型身份的准确性。根据证据 4 的上述记载可知，证据 4 是在单幅图像处理过程中对输入图像进行了表情去除以及设置随机表情的处理，从而实现身份与表情的解耦，而本专利权利要求 1 需要利用参考图像的目标表情系数作为模型输出结果并用于对模型进行调整，可见证据 4 与本专利采用了相反的图像处理方法，并且证据 4 仅公开了一种利用单幅图像进行 3D 人脸重建的方法，并不涉及音频相关的特征，也未给出需要对训练后的网络模型进行调整的启示，正是采用了上述区别技术特征，权利要求 1 获得了生成目标动态图像中的图像更贴近待处理图像中对象的有益的技术效果，因此，权利要求 1 相对于证据 3、证据 4 的结合具有突出的实质性特点和显著的进步，具备专利法第 22 条第 3 款有关创造性的规定。

另外，根据上述对证据 3 中各个部分的比对分析可知，证据 3 的各个部分实际涉及图像生成模型的各个阶段，并不是可以相互组合的实施例，因此请求人关于权利要求 1 相对于证据 3 的各个实施例的结合不具备专利法第 22 条第 3 款规定的创造性的无效理由不成立。

（4.1.3）以证据 5 作为最接近的现有技术

证据 5 公开了（参见说明书第 40-187 段及图 2-6）图像生成模型的训练方法，如图 2 所示，包括以下步骤：

S202, 从样本图像集合中获取与当前样本对象对应的当前样本图像, 其中, 样本图像集合中包括与多个样本对象分别对应的多个样本图像子集, 一个样本图像子集中包括对应的一个样本对象的多个样本图像;

S204, 利用处于训练状态的图像生成模型中的属性特征提取网络提取当前样本图像的属性特征向量, 并获取与前样本图像中的当前样本对象的对象标识对应的标识特征向量, 其中, 属性特征向量用于指示当前样本图像中的当前样本对象的三维图像特征;

S206, 将属性特征向量和标识特征向量输入训练状态的图像生成模型中的图像生成网络, 得到参考样本图像;

S208, 利用处于训练状态的图像生成模型中的判别网络获取参考样本图像与当前样本图像的相似度匹配结果;

S210, 在基于相似度匹配结果确定的训练损失满足训练条件的情况下, 将处于训练状态的图像生成模型中的属性特征提取网络和图像生成网络作为目标图像生成模型中的目标属性特征提取网络和目标图像生成网络, 其中, 目标图像生成模型用于基于待重建图像生成目标重建图像, 并基于驱动参数驱动所述目标重建图像中的目标对象。

图 3、图 4、图 5 对本申请上述实施方式的一个完整训练过程进行说明。在本实施方式中, 提出一种基于生成式的高精度人像重建和驱动方法, 主要包括三个网络: 一个属性提取网络 (Attribute Encoder), 用于提取图像中人物的姿态、表情、光照等信息; 一个身份提取网络 (Identity Encoder), 用于提取人物的身份信息; 以及一个图像生成器 (Generator), 用于生成特定人物在指定的姿态、表情、光照下的图像。如图 3 所示, 在数据准备阶段, 可以从获取同人图片和视频 (视频抽帧得到图片), 对数据进行整理, 每个人给定一个编号 ID_i , 一共有 N 个人, 每个人有若干图片, 姿态表情光照等属性各不相同, 每张图片能索引到对应的人物编号 ID_i 。

作为一种可选的实施方式, 如图 5 所示, 上述图像生成方法可以包括以下步骤:

S502, 获取目标图像生成模型, 其中, 目标图像生成模型为根据样本图像集合对处于训练状态的图像生成模型经过对抗生成训练得到的图像生成模型, 样本图像集合中包括与多个样本对象分别对应的多个样本图像子集, 一个样本图像子集中包括对应的一个样本对象的多个样本图像;

S504, 利用目标图像生成模型中的目标属性特征提取网络提取待重建图像的第一属性特征向量, 并获取待重建图像的身份特征向量, 其中, 属性特征向量用于指示待重建图像中的图像对象的三维图像特征;

S506, 将第一属性特征向量和身份特征向量输入目标图像生成模型中的目标图像生成网络, 得到与待重建图像匹配的目标重建图像, 其中, 目标重建图像中包括目标对象, 目标对象为待重建图像中的图像对象。

可以理解的是, 在本申请上述实施方式中, 可以通过待重建图像本身的第一属性特征向量和身份特征向量对带重建图像进行重建操作, 得到目标重建图像。可以理解的是, 在目标重建图像中包括的目标对象为待重建图像中的图像对象, 且目标对象执行的驱动动作与待重建图像中的图像对象的动作相同。也就是说, 在

本申请上述实施方式中，执行的是基于原图重建的操作。

在一种可选的实施方式中，上述将第一属性特征向量和身份特征向量输入目标图像生成模型中的目标图像生成网络，得到与待重建图像匹配的目标重建图像之后，还包括：

S1，获取驱动参数；

S2，将驱动参数和身份特征向量输入目标图像生成模型中的目标图像生成网络，得到与目标重建图像匹配的目标驱动图像，其中，目标驱动图像中包括执行目标驱动动作的目标对象，目标驱动动作为与驱动图像中的图像对象的对象动作相匹配的动作，目标对象为待重建图像中的图像对象。

可选地，上述获取驱动参数包括以下之一：

方式一、获取驱动图像，并利用目标图像生成模型中的目标属性特征提取网络提取驱动图像的第二属性特征向量，将第二属性特征向量作为驱动参数；

方式二、获取为目标重建图像配置的驱动参数。

可以理解的是，在本实施方式中，提供了两种获取驱动参数的方法，第一种方式可以根据驱动图像获取的第二属性特征向量作为驱动参数；第二种方式可以是直接获取用于驱动目标对象的驱动参数，上述驱动参数可以是根据驱动指令生成，例如，在驱动指令为“作出一个微笑的表情”的情况下，对应的驱动参数可以是用于控制上述目标对象作出上述微笑表情的特征参数。

在一种可选的实施方式中，上述获取待重建图像的身份特征向量包括以下之一：

方式一、获取目标对象标识，将目标对象标识的标识特征向量作为身份特征向量，其中，目标对象标识用于指示待重建图像中的图像对象；

方式二、获取待重建图像，并根据目标图像生成模型中的目标身份特征提取网络提取待重建图像的身份特征向量。

可以理解的是，在基于目标图像生成网络进行图像重建和驱动的过程中，可以至少包括两种使用方式。

在第一种使用方式中，由于在图像生成模型的训练过程中已经对样本图像中的样本对象的对象身份特征进行了充分学习，进而可以在仅获取到驱动图像的情况下，根据用户选定的目标对象标识，获取与对象标识匹配的标识特征向量（即直接获取训练过程中所使用的样本对象的身份特征），并基于驱动图像中的图像对象的属性特征对上述目标对象标识对应的目标对象进行驱动显示，以显示目标对象标识对应的图像对象执行与驱动图像中的图像对象相匹配的对象动作或对象表情；

在第二种使用方式中，以在获取到待重建图像和驱动图像的情况下，利用训练好的目标身份特征提取网络提取待重建图像的身份特征向量，并基于驱动图像中的图像对象的属性特征对上述待重建图像的图像对象进行驱动显示，以显示待重建图像中的图像对象执行与驱动图像中的图像对象相匹配的对象动作或对象表情。

以下对上述图像生成模型的几个具体应用场景进行说明。在一种可选的场景中，可以在用户进行游戏任务的过程中，在用户授权许可的情况下，对当前用户当前面部的图像进行采集，并将采集到的图像作为驱动

图像输入上述图像生成模型，再根据用户选择的虚拟形象作为待重建图像，进而基于用户的当前表情驱动游戏中的虚拟形象作出对应的虚拟表情；

在另一种可选的场景中，可以在用户进行直播操作的过程中，在用户授权许可的情况下，对当前用户当前面部的图像进行采集，作为待重建图像；并获取用户输入的音频的音频特征，基于音频特征生成对应的驱动图像，并将驱动图像和待重建图像输入上述图像生成模型，进而基于用户选择的背景音乐所生成的驱动图像中的对象动作，驱动基于用户的当前面部图像进行虚拟直播，例如可以是虚拟唱歌直播。

上述三个网络训练好以后，对于任意给定的真实图像，我们就可以提取他的身份特征和属性特征，然后通过 mapping 网络将图像对应的属性特征和身份特征映射成 StyleGAN 的 latent 向量给到生成器就可以生成重建后的图像。重建完成后，就可以进行驱动，可以从驱动图像上使用属性提取网络抽取姿态表情光照等属性信息，然后和上述重建过程中得到的身份特征进行组合，同步输入至生成器（图像生成子网络），就可以生成驱动后的图像。如图 6 所示，为基于本申请上述实施方式重建并驱动生成的图像序列的示意图。如图 6 中的第一行第一张图，可以是用户输入的待重建图像（即真实图像），图 6 中的第一行第二张至第九张图，为一组驱动图像序列。

根据证据 5 的上述记载可知，证据 5 的图 2 及其相关文字描述了一种图像生成模型的训练方法，图 3、图 4、图 5 及其相关文字描述了一个完整训练过程，其属于模型训练阶段；图 6 及其相关文字描述了利用图 2-5 所示的训练好的图像生成模型对一个真实图像进行原图生成，并利用采集到的图像作为驱动图像输入上述图像生成模型进行同步复制用户动作和表情，其属于模型应用阶段。根据上述记载可知，证据 5 中对于真实图像的处理，首先需要提取真实图像的身份特征，根据提取的身份特征去样本图像集合中获得对应该身份特征的对象分别对应的多个样本图像子集，一个样本图像子集中包括对应的一个样本对象的多个样本图像，然后据此对真实图像进行原图重现，再根据驱动信息进行驱动图像，也就是说真实图像中的对象是该图像生成模型的模型训练过程中的样本图像集合中的样本对象，这个样本图像集合中有该真实图像中的对象的多个图像，利用这些图像完成原图生成，生成的图像与原图的相似度是符合损失函数要求的。即由于证据 5 在训练过程中的样本图像集合中已经包括真实图像中的对象，且生成图像的模型输出结果在前面验证后已经满足要求。也就是说证据 5 中公开了图像生成模型的训练方法以及对训练完成的图像生成模型的应用，其图像生成模型经过训练后可以直接用于图像生成，并不涉及也不需要训练后图像生成模型的再调整的内容，更不涉及具体的调整手段；而本专利权利要求 1 首先有训练后的生成网络模型，然后再利用用户输入的图像作为参考图像对训练后的生成网络模型再进行调整，并且这次对于模型的调整是基于输入的参考图像，该参考图像的对象与待处理图像的对象相同，并不涉及音频内容，以及权利要求 1 还记载了具体的模型调整手段。可由此可知，本专利权利要求 1 与证据 5 对于模型的训练方法是不同的。

因此，本专利权利要求 1 与证据 5 的区别技术特征在于：基于所述参考图像和训练后的生成网络模型，确定目标头部动作特征和目标表情系数特征；所述训练后的生成网络模型用于基于输入的所述参考图像生成

多个预测图像，并基于各所述预测图像和所述参考图像之间的区别，确定目标头部动作特征和目标表情系数特征；基于所述目标头部动作特征和所述目标表情系数特征对所述训练后的生成网络模型进行调整，得到目标生成网络模型；基于所述参考音频、所述参考图像和所述目标生成网络模型，对待处理图像进行处理，得到目标动态图像；所述待处理图像与所述参考图像中的图像对象相同。基于上述区别技术特征可以确定权利要求 1 实际解决的技术问题是如何降低生成精准图像的成本。

根据上述对证据 5 中各个部分的比对分析可知，证据 5 的各个部分实际涉及图像生成模型的各个阶段，并不是可以相互组合的实施例，而且证据 5 并未公开可以对训练后的图像生成模型的进行再调整的内容，也不涉及具体的调整手段，正是基于上述区别技术特征，权利要求 1 获得了能够降低生成精准图像成本的有益的技术效果。因此，请求人关于权利要求 1 相对于证据 5 的各个实施例的结合不具备专利法第 22 条第 3 款规定的创造性的无效理由不成立。

（4.2）关于权利要求 2-12

权利要求 10 保护与权利要求 1 的方法权利要求对应的装置权利要求，基于与权利要求 1 相同的评述，权利要求 10 相对于上述相应证据组合也具备专利法第 22 条第 3 款规定的创造性。

由于请求人关于权利要求 1 相对于上述相应证据不具备创造性的无效理由不成立，基于引用基础，请求人关于从属权利要求 2-9、权利要求 11-12 相对于相应证据不具备专利法第 22 条第 3 款有关创造性的无效理由也不成立。

综上所述，请求人所有的无效理由均不成立。合议组现依法作出如下决定。

三、决定

维持 202410022841.6 号发明专利权有效。

当事人对本决定不服的，根据专利法第 46 条第 2 款的规定，可以自收到本决定之日起三个月内向北京知识产权法院起诉。根据该款的规定，一方当事人起诉后，另一方当事人作为第三人参加诉讼。

合议组组长：熊婷
主审员：吴士芬
参审员：林甦

